논문 2006-43CI-6-5

Monk's Problem에 관한 가우시안 RBF 모델의 성능 고찰

(A Performance Study of Gaussian Radial Basis Function Model for the Monk's Problems)

신 미 영*, 박 준 구*

(Miyoung Shin and Joon Goo Park)

요 약

데이터 마이닝(data mining)이란 대량의 데이터에 내재되어 있는 숨겨진 패턴을 찾아내기 위한 분석 기술로서 지금까지 많은 연구가 진행되어 왔지만, 현재의 데이터 마이닝 연구는 ad-hoc 문제와 같은 해결되어야 할 중요한 이슈들이 있다. 즉, 개별적 문제에 대해 설계된 마이닝 기법이 주로 사용되는 까닭에 여러 문제에 통합적으로 적용될 수 있는 시스템적 마이닝 기법에 관한 연구가 요구되고 있다. 본 논문에서는 이러한 핵심 데이터 마이닝 태스크 중의 하나인 분류 모델링 방법으로 방사형 기저 함수(radial basis function, RBF) 모델의 성능을 고찰하고 그 유용성(usefulness)을 살펴보고자 한다. 특히, 대표적인 마이닝 관련 벤치마킹 데이터인 Monk's problem 분석을 위해 *RC(Representation Capacity)* 기반 알고리즘을 사용하여 RBF 모델을 구축하고 분류 성능을 기존의 연구 결과와 비교 고찰한다. 그리하여 RBF 모델의 분류 성능 면에서의 우수성 뿐만 아니라 모델링 과정을 체계적인 방식으로 적절히 제어할 수 있음을 보여주고, 이를 통해 현재의 ad-hoc 방식의 문제를 어느 정도 해결할 수 있음을 보여준다.

Abstract

As an analytic method to uncover interesting patterns hidden under a large volume of data, data mining research has been actively done so far in various fields. However, current state-of-the-arts in data mining research have several challenging problems such as being too ad-hoc. The existing techniques are mostly the ones designed for individual problems, so there is no unifying theory applicable for more general data mining problems. In this paper, we address the problem of classification, which is one of significant data mining tasks. Specifically, our objective is to evaluate radial basis function (RBF) model for classification tasks and investigate its usefulness. For evaluation, we analyze the popular Monk's problems which are well-known datasets in data mining research. First, we develop RBF models by using the representational capacity based learning algorithm, and then perform a comparative assessment of the results with other models generated by the existing techiques. Through a variety of experiments, it is empirically shown that the RBF model has not only the superior performance on the Monk's problems but also its modeling process can be controlled in a systematic way, so the RBF model with RC-based algorithm might be a good candidate to handle the current ad-hoc problem.

Keywords: data mining, classification model, radial basis functions, Monk's problems, performance study

I. 서 론

데이터 마이닝(data mining)이란 대량의 데이터에 내재되어 있는 숨겨진 패턴을 찾아내기 위한 분석 기술로서, 웹 로그 데이터 분석에서부터 Telemedicine을 위한 오디오/비디오 분석이나 RFID 데이터 분석, DNA 마이크로어레이 등의 바이오 의료 데이터 분석 등에 이르기까지 여러 분야에서 활발한 연구가 진행되어오고 있다. 특히, 다양한 응용 분야에서 요구되는 핵심 데이터 마이닝 태스크 중의 하나인 분류 문제(classification problem)는 선분류된(pre-classified) 예제(instance)들을 사용하여 새로운 예제들의 해당 클래스를 설정하는 작업을 말한다. 이러한 분류 모델링을 위해 CART[1], CHAID[2] 등의 통계적인 접근 방식들이 존재하지만, 이 경우 데이터의 확률 분포에 대한 가정에 기반하고 있기 때문에 실제 데이터에 적용하는 데에는 많은 제약이 있

* 정회원, 경북대학교 전자전기컴퓨터학부
(School of Electrical Engineering and Computer Science, Kyungpook National University.)
접수일자: 2006년7월14일, 수정완료일: 2006년10월30일

다. 한편, ID3, C4.5[3]와 같은 결정 트리 방식은 해석 가능한 규칙을 제공할 수 있다는 장점이 있지만, 결과의 일반화(generalization)를 위해서는 휴리스틱 방식에 기반한 가지치기(pruning), 속성값의 그룹화 등과 같은 상당한 계산 과정을 거쳐야 한다. 또한, 신경망(neural networks)은 적용을 위한 제약조건이 없고 일반화 능력이나 복잡한 패턴에 대한 학습 능력이 뛰어난 장점이 있지만, ad-hoc 방식의 학습 과정과 결과 해석이 어려워 신경망 모델을 적용하는 데에 많은 어려움이 있다. 최근 Yang et al.[4]이 지적한 바와 같이, 데이터 마이닝 연구는 ad-hoc 방식의 모델링 문제와 같은 해결해야 할 중요한 이슈들이 있다. 즉, 개별적인 문제에 대해 설계된 마이닝 기법을 대신하여 여러 문제에 통합적으로 적용될 수 있는 체계적 마이닝 기법에 관한 연구가 요구되고 있다.

본 논문에서는 데이터 마이닝 응용을 위한 분류 문제의 모델링 방법의 하나로서 신경망의 하나인 방사형 기저 함수(radial basis function, RBF) 모델의 성능을 고찰하고 그 유용성(usefulness)을 살펴보는 데에 목적이 있다. 특히, 대표적인 마이닝 관련 벤치마킹 데이터인 Monk's problem[5]분석을 위해 *RC(Representation Capacity)* 기반 알고리즘을 사용하여 RBF 모델을 구축하고 분류 성능을 기존의 연구 결과와 비교 고찰하고자 한다.

본 논문의 구성은 다음과 같다. 먼저 제 II장에서는 Monk's problems의 문제 설명 및 기존 연구 결과에 대해 살펴본다. 제 III장에서는 RBF 모델에 대해 간략히 설명하고, RC 기반 RBF 학습 알고리즘 및 데이터 마이닝에의 응용을 위한 모델링 이슈를 기술한다. 제 IV장에서는 Monk's problems에 관한 실험 진행 방법 및 결과를 제시하고, 분류 성능의 고찰 및 모델 선택에 따른 trade-off를 살펴본다. 제 V장에서는 논문을 요약하고, 결론을 맺는다.

II. Monk's Problems

1. 문제의 설명

Monk's problems[5]은 인위적으로 생성된 선분류된(pre-classified) 로봇 데이터로서 다른 복잡도를 가진 총 세 가지 문제로 구성되어 있으며, 각 학습(training) 샘플은 로봇을 묘사하는 여섯 개의 이산치 속성(discrete-valued attributes)에 의해 표현된다. 각 문제는 속성(attributes)들의 논리적 관계 조합에 의해 표현되는 특정 로봇 클래스를 학습하기 위한 것으로, 각 로봇 샘플은 이러한 클래스에 속하거나 속하지 않을 수 있다. 아래 표 1은 Monk's problem에서 사용하는 로봇 관련 여섯 개의 속성과 이 속성들이 취할 수 있는 값의 범위를 보여주고 있다.

표 1. Monk's problems에 대한 속성들과 각 속성이 취할 수 있는 값의 범위

Table 1. Attributes and their values for THE Monk's problems.

속성		값의 범위
X_1	head_shape	round, square,octagon
X_2	body_shape	round, square, octagon
X_3	is_smiling	yes, no
X_4	holding	sword, balloon, flag
X_5	jacket_color	red, yellow, green, blue
X_6	has_tie	yes, no

표 1에 나타난 각 속성은 취할 수 있는 값의 개수만큼의 길이를 가진 이진 문자열로 인코딩할 수 있으며, 여섯 개의 속성들은 총 17개 (3+3+2+3+4+2 = 17)의 이진 문자열로 표현할 수 있다. 또한, 432개의 서로 다른 속성 조합(3*3*2*3*4*2 = 432)이 가능하다. 학습 데이터는 이러한 432개 속성 조합의 부분집합으로 구성되며, 이러한 학습 데이터로부터 분류 모델을 생성하고, 전체 데이터에 관해 일반화 성능(generalization performance)을 평가할 수 있다. 그리하여, 아래의 모든 문제에 대하여, 전체 432개 샘플이 테스트 데이터로 사용된다. 세 가지 Monk's problems은 아래와 같다[4,5].

▷ [Monk's 1] *(head_shape = body_shape) 또는 (jacket_color = red)인 로봇 클래스를 분류*

알고리즘이 단순한 개념을 학습할 수 있는지를 확인하기 위하여 설계된 상대적으로 간단한 테스트 문제이다. 432개의 가능한 샘플로부터 124개를 무작위로 선택하여 학습 데이터가 구성되었으며, 노이즈가 포함되지 않았다.

▷ [Monk's 2] *정확히 두 개의 속성(attributes)에 대해서 첫번째 속성값을 가지는 로봇 클래스를 분류*

Monk's 1에 비해 훨씬 복잡한 관계를 나타내는 문제로서 학습하기 매우 어려운 문제로 알려져 있다. 432개의 가능한 샘플로부터 169개를 무작위로 선택하여 학습 데이터가 구성되었다. 이 경우에도 노이즈가 포함되지 않았다.

▷ [Monk's 3] *{(jacket_color = green) and (holding = sword)} 또는 {(jacket_color ≠ blue) and*

(body_shape ≠ octagon)}인 로봇 클래스를 분류

복잡도 면에서는 Monk's 1과 비슷하지만 약 5%의 노이즈를 학습 데이터에 포함하고 있다는 특징이 있다. 432개의 가능한 샘플로부터 122개를 무작위로 선택하여 학습 데이터가 구성되었고, 이 중 5%에 해당하는 6개의 오분류된 샘플을 학습 데이터에 포함하고 있다.

2. 관련 연구

본래 Monk's problems은 Second European Summer School on Machine Learning에서 여러 학습 기술을 비교하기 위해 최초로 사용하였으며, 최근까지 많은 데이터 마이닝 연구자들이 다양한 분류 알고리즘의 우수성을 테스트하기 위한 중요한 벤치마킹 데이터로서 활용하고 있다[6,7,8,9,10,11]. 종종 본래의 연구에서 사용된 것과 다른 데이터 분포를 사용하는 경우도 있기 때문에 수치상의 성능만으로 알고리즘의 우수성을 판단하기에는 다소 어려움이 있지만, 현재까지의 Monk's problems에 대한 전반적인 분류 성능의 수준을 이해하기 위하여 아래와 같이 최근 연구 결과물들의 성능을 살펴보았다. 표 2에서는 최근 발표된 알고리즘들의 Monk's problems에 관한 분류 성능을 요약 제시한다.

Xiong et al.[6]은 경험적인 특징 공간(empirical feature space)에서 데이터의 클래스 분별력(class separability)을 극대화하도록 커널을 최적화하고 이를 Monk's problems 분석에 적용한 바 있다. 데이터를 정규화하고 동일한 크기의 세 부분으로 나누어 이 중 한 부분을 경험적 데이터로, 나머지 두 부분을 학습 데이터와 테스트 데이터로 각각 사용한 결과, Monks1, Monks2, Monks3에 대하여 대략 85%, 78%, 96%의 분류 정확도를 나타내었다.

한편, Mitchell's 논문[7]에서는 TRACA 시스템의 Monk's problem에 대한 분석 결과로서 각각 97%, 70%, 93%의 분류 정확도를 보여주고 있다. Casey et al.[8]는 일반화 성능이 멀티 네트워크 상에서의 동시 학습을 통해 얼마나 개선될 수 있는 지를 살펴보기 위하여 in-situ learning을 적용한 결과 세 가지 Monk's problem에 대하여 대략 90%, 70%, 79%의 분류 정확도를 얻었다. Huang et al.[10]은 불필요한 데이터 특징을 제거하기 위해 차원 감소 기법을 적용하였고 이를 통해 Monks1과 Monks3에 관한 성능 개선 결과를 보여주고 있다. 반면, Monks2에 관한 차원 감소는 이루어지지 않았다. Saxon et al.[11]는 XCS를 Monk's problems에 적용한 결과를 제시하고 있으며 반복수가 증가함에 따라 분류 성능이 개선됨을 보여주고 있다. Toh et al.[9]는 패턴분류를 위하여 축소된 다변량 다항식(reduced multivariate polynomial) 모델을 제안하고 10-fold cross vadliation의 평균결과로서 Monks1, Monks2, Monks3 각각에 대하여 대략 76%, 65%, 90%의 분류 정확도 값을 얻은 바 있다.

표 2. Monk's problems에 대한 최근 알고리즘들의 테스트 데이터에 관한 분류 성능 요약

Table 2. A summary of classification performance of recent algorithms on test data for THE Monk's problems.

적용 방법	Monk's 1	Monk's 2	Monk'3
Xiong et al. (2005)[6]	85%	78%	96%
Mitchell (2003)[7]	97%	70%	93%
Casey et al. (2004)[8]	90%	70%	79%
Toh et al. (2005)[9]	76%	65%	90%

III. 가우시안 RBF 모델

1. 모델 설명

RBF 모델은 몇 가지 독특한 특징과 수학적 특성을 지닌 신경망 모델의 한 형태이다. 입력층, 출력층, 그리고 하나의 은닉층으로 이루어져 있으며, 모델 변수에 비선형성(non-linearity)과 선형성(linearity)을 함께 나타내는 것이 가능하고, 이러한 변수들을 모델링 과정에서 분리하여 다룰 수 있다. 더욱이 가우시안 RBF 모델은 최적 근사(best approximation)와 보편 근사(universal approximation)라는 중요한 수학적 성질을 가지고 있다.

그림 1은 입출력 데이터 x_i와 y에 대하여 은닉층에 m개의 커널을 가진 일반적인 RBF 모델의 구조를 보여주고 있다. 입력층은 n개의 d차원의 입력 데이터 $\{x_i \in R_d, i = 1,...,n\}$ 각각에 대하여 은닉층에 있는 k개의 커널을 기반으로 데이터 변환하여 커널 출력값 $\phi_1(x_i), \phi_2(x_i), ..., \phi_k(x_i)$ 을 생성한다. 여기서 $\phi_j(x_i)$은 x_i에 대한 j번째 커널의 출력값을 나타내며, 이것은 가중치 w_j를 곱한 후 더해져 최종 모델 출력값 $y = \sum_{j=1}^{k} w_j \phi_j(x_i)$을 생성한다. 그리하여, 입력 데이터에 대해 RBF 모델은 궁극적으로 다음과 같이 n×m 크기의 디자인 행렬 $\mathbf{\Phi}$로 변환된다.

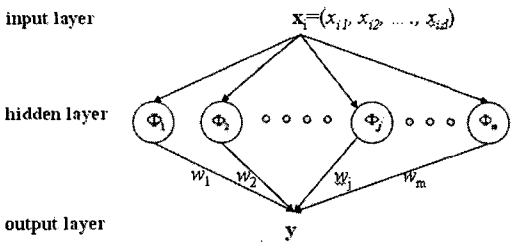


그림 1. RBF 모델 구조
Fig. 1. RBF model structure.

$$\Phi = \begin{bmatrix} \Phi_1(x_1) & \Phi_2(x_1) & \cdots & \Phi_m(x_1) \\ \Phi_1(x_2) & \Phi_2(x_2) & \cdots & \Phi_m(x_2) \\ \vdots & \vdots & & \vdots \\ \Phi_1(x_n) & \Phi_2(x_n) & \cdots & \Phi_m(x_n) \end{bmatrix} \quad (1)$$

수식 (1)에서 $\Phi_j(\cdot)$ 는 가우시안 커널이 주로 사용되며, $\phi_j(x) = \exp(-\|x-\mu_j\|^2/2\sigma_j^2)$이 된다. 여기서 μ_i 와 σ_i 은 커널 변수로서 가우시안 커널의 중심(center)과 폭(width)을 나타낸다. 그리하여, 가우시안 커널 RBF 모델의 최종 출력값 $\hat{y}$은 아래식 (2)와 같다. 이 때, $\|\cdot\|$은 유클리디안 거리를 나타낸다.

$$\hat{y} = f(\mathbf{x}) = \sum_{j=1}^{k} w_j \exp\left(-\frac{\|\mathbf{x}-\boldsymbol{\mu}_j\|^2}{2\sigma^2}\right) \quad (2)$$

결과적으로, 가우시안 커널을 가진 RBF 모델은 은닉층을 통해 데이터의 비선형적 변환을 정의하는 데 필요한 커널의 개수(m) 및 커널의 중심(μ)과 폭(σ), 그리고 은닉층으로부터의 결과를 출력층에 선형적으로 대응시키는 역할을 하는 커널 출력값의 가중치(w)에 의해 정의된다. 즉, RBF 모델링 문제는 주어진 데이터 D={(X, y)}에 대하여, 변수 집합 P=(m,μ,σ,w)의 적절한 값을 학습하는 문제로 요약될 수 있다.

2. *RC* 기반 RBF 학습 알고리즘

RBF 모델링 시에 고려해야 할 중요한 사항 중의 하나는 학습 데이터에 대한 모델의 복잡도(model complexity)를 결정하는 문제이다. 무엇보다 overfitting과 underfitting 사이에서 적절한 균형을 유지하는 최적의 복잡도를 찾아내어야 한다. 일반적으로 너무 단순한 모델은 underfitting을 나타내고 너무 복잡한 모델은 overfitting을 나타내는 경향이 있기 때문에 두 경우 모두 새로운 데이터에 대해 좋지 않은 성능을 보인다.

본 논문에서는 RBF 모델 학습을 위해 *RC* 기반 알고리즘을 사용함으로써 RBF 모델의 복잡도에 따른 양극단 사이에서 적절한 균형을 유지하도록 모델 변수값을 자동 설정하도록 한다. *RC* 기반 RBF 학습 알고리즘[12]은 모델 변수를 보다 체계적인 방법으로 찾아내기 위하여 *Representational Capacity* (*RC*) 개념을 도입하고 이에 따라 모델 복잡도 및 변수를 최적화하는 알고리즘이다. 사용자에 의해 명시된 *RC* 기준치를 만족하는 최적화된 RBF 모델을 보다 체계적인 방법으로 찾아낼 수 있다. 상기 알고리즘을 간략히 요약하면 다음과 같다.

INPUT:
X : A data matrix of size $n \times d$
δ : RC criterion taken as $0 < \delta < 1$.
Empirically, we chose δ=0.01, 0.001, and 0.0001.
σ : A global width chosen as in $0 < \sigma < \sqrt{d}$.

ALGORITHM
1. Construct an interpolation matrix $\mathbf{D} = [d_{ij}]_{i=1,\ldots,n, j=1,\ldots,n}$
   - $d_{ij} = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2\sigma^2)$
2. Determine the number of kernels (m).
   - Compute singular value decomposition (SVD) such that $\mathbf{U}^T\mathbf{D}\mathbf{V} = diag(s_1, s_2, \cdots, s_m)$
   - m = rank($\mathbf{D}$, $s_1 \times \delta$)
3. Determine kernel parameters ($\mu_1, \mu_2, \ldots, \mu_m$).
   - Conduct QR factorization with column pivoting such that $\mathbf{Q}^T\mathbf{V}_{n\times m}^T\mathbf{P} = R_{n\times n}$
   where $\mathbf{V}_{n\times m}$ = upper left $n \times m$ elements of V.
   - Compute $\mathbf{X}^T\mathbf{P} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n]$
   - $(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_m) \equiv (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_m)$
4. Build a design matrix $\boldsymbol{\Phi} = [\phi_{ij}]_{i=1,\ldots,n, j=1,\ldots,m}$
   - $\phi_{ij} = \exp(-\|\mathbf{x}_i - \boldsymbol{\mu}_j\|^2/2\sigma^2)$
5. Compute weights $w \equiv (w_1, w_2, \ldots, w_m)$
   - $w = \Phi^+ y$

보다 상세한 내용은 참고문헌[12]에서 참조할 수 있다.

3.3. 데이터 마이닝 응용을 위한 RBF 모델링 이슈

앞서 기술한 바와 같이, 최근 데이터 마이닝 알고리즘이 직면한 가장 중요한 문제 중의 하나는 *ad-hoc* 방식의 학습 과정이며, 여러 문제에 체계적 방식으로 적용될 수 있는 통합된 이론적 방식에 관한 연구가 요구되고 있다[1]. 이러한 관점에서 볼 때, 3.2절에서 기술한 *RC* 기반 RBF 모델 학습 알고리즘을 데이터 마이닝 문제에 적용하기 위해서는 *RC* 기준치를 조절하는 δ값을 어떻게 결정할 것인가에 대한 문제가 있다.

변수 δ는 RBF 모델의 복잡도를 조절하는 제어 변수이며, $0 < \delta < 1$ 사이의 값을 가진다. δ=0인 경우, 모델의 데이터 표현 능력인 *RC*(=1-δ) 값이 1에 대응하고, 이것은 학습 데이터를 완전하게 나타내는 모델의 복잡

도를 의미한다. 한편, δ의 값이 커질수록, 즉 *RC* 값이 1에서 멀어질수록, 학습 데이터에 부합(fitting)되는 정도가 조금씩 낮아지는 모델 복잡도를 찾는다는 것을 의미한다. 따라서, 입력변수 δ 값을 설정할 때, 데이터의 노이즈 정도나 문제의 복잡도에 대한 판단에 따라 0에 근접한 값부터 점차 증가된 값으로 시도해 볼 수 있다. 또한, 성능에 비해 복잡도가 낮은 모델을 선호할 경우 상대적으로 큰 δ 값을 사용하고, 복잡도에 상관없이 성능이 우선될 경우, 작은 δ 값을 사용함으로써 실제 상황을 반영하여 학습 과정을 진행해 갈 수 있다. 그리하여 최적의 모델 복잡도를 성능과 모델 복잡도의 타협점(trade-off)을 고려하여 적절하게 조절할 수 있다.

IV. Monk's Problem에 대한 실험 결과

1. 실험 데이터

본 논문에서 사용한 실험 데이터는 각 데이터 샘플이 여섯 개의 속성으로 구성되며[5], 각 속성들은 길이 k개의 이진 벡터(binary vector) 로 나타내어진다. 여기서, k는 각 속성이 취할 수 있는 값의 종류의 수와 같다. 즉, 표 1에서와 같이, 속성 X_1 (head_shape)은 세 가지의 다른 값 (round, square, octagonal)을 취할 수 있기 때문에, 길이 3의 이진 벡터로서 나타낼 수 있다. 만약 head_shape이 round이라면, 벡터의 첫 번째 값은 1이고, 두 번째 및 세 번째 값은 0이 된다.다른 속성에 대해서도 이와 유사한 형태로 표현된다. 따라서, 이러한 변환 체계를 사용하면 각 입력 데이터 샘플은 17차원(3+3+2+3+4+2)의 벡터로 변환되고, 출력값은 이진 변수로서 표현된다.

2. 실험 진행 방법

본 논문에서의 실험은 각 문제별로 주어진 학습 데이터에 대하여 다음과 같이 진행되었다. 먼저, 커널 변수 (m,σ)의 전체 공간에서 각 (m,σ) 조합에 대하여 RBF 모델을 생성하고 이를 기반으로 테스트 오차 표면(test error surface) 및 컨투어(contours)을 시뮬레이션 하였다. 테스트 오차 표면 및 컨투어는 커널 변수 (m,σ) 값의 선택에 따라 결과 모델의 테스트 오차가 변화되는 전체적인 추이를 나타내며, Monks1, Monks2, Monks3에 대한 테스트 오차 표면 및 컨투어는 각각 그림 2, 4, 6 및 그림 3, 5, 7과 같다.

둘째, 앞의 제 3.2절에서 기술된 알고리즘을 사용하여 다양한 *RC* 기준치(즉, δ=0.01, 0.001, 0.0001)에 따라 가우시안 커널 RBF 모델을 생성하고, 테스트 데이터에 대한 오분류율로서 테스트 오차를 계산한다. 이 때, *RC* 기준치의 선택에 따라 결과 모델이 어떠한 차이를 나타내는지와 시뮬레이션된 전체 테스트 오차 표면 상에서 이러한 모델이 가지는 의미를 살펴본다.

4.3 Monks1에 대한 실험 결과

124개의 데이터 샘플로 구성된 Monks1 학습 데이터에 대하여 m=(1:6:124)와 σ=(0.2:0.2:3)의 범위에서 모든 가능한 (m,σ) 조합에 대한 RBF 모델을 생성하고 이에 대한 테스트 오차 표면 및 컨투어를 그림 2, 3과 같이 각각 생성하였다. 그림 3에서 흰색으로 나타나는 공간은 테스트 오차가 최적인 범위*를 나타낸다.

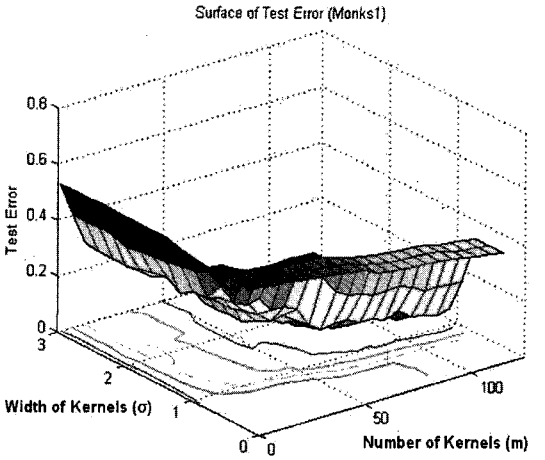


그림 2. Monks1에 대한 전체 커널 변수(m,σ) 공간에서 생성된 테스트 오차 표면

Fig. 2. Surfaces of test errors generated over kernel parameter (m,σ) space for Monks1.

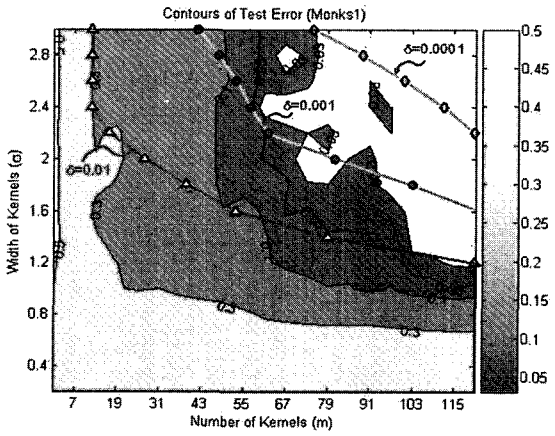


그림 3. Monks1에 대한 테스트 오차 컨투어 및 δ별 탐색 후보 모델의 예

Fig. 3. Contours of test errors and candidate models explored for different δ in Monks1.

* Monk1에 대하여 최근까지 발표된 결과 중 테스트 오차 최저치인 0.03 (즉, 오분류율이 3%)보다 적은 범위를 나타낸다.

표 3. Monks1에서 δ값별 최적의 영역에 속하는 테스트 오차를 가지는 모델들

Table 3. Selected models having the test errors within the optimal range for different δ in Monks1.

δ	σ	*m*	TestCE	Classification Accuracy
0.01	1.2	121	0.025	97.5%
0.001	1.2	124	0.027	97.3%
	1.4	107	0.023	97.7%
0.0001	1.4	124	0.014	98.6%
	1.6	124	0.016	98.4%
	1.8	124	0.012	98.8%
	2.0	124	0.009	99.1%
	2.2	121	0.009	99.1%
	2.4	112	0.019	98.1%
	2.6	101	0.023	97.7%
	2.8	89	0.021	97.9%

앞의 제 3.2절에 기술된 RC 기반 학습 알고리즘을 사용하여 RBF 모델을 생성하기 위하여, RC 기준치로서 δ =0.01, 0.001, 0.0001의 세 가지 값을 사용하였고, 각 δ값에 대하여 알고리즘에 의해 자동 탐색된 모델들은 그림 3에 표기된 바와 같다. 구체적으로 δ=0.01인 경우, (m, σ)={(12,3), (12,2.8), (12,2.6), (12,2.4), (17,2.2), (27,2), (39,1.8), (53,1.6), (79,1.4), (121,1.2), (124,1)}에 대응하는 모델들이 탐색되었으며, δ=0.001인 경우에는 (m, σ)={(42,3), (48,2.8), (53,2.6), (57,2.4), (62,2.2), (81,2), (103,1.8), (122,1.6), (124,1.4), (124,1.2), (124,1)}이 탐색되었다. δ=0.0001인 경우에는 (m,σ)={(75,3), (89,2.8), (101,2.6), (112,2.4), (121,2.2), (124,2), (124,1.8), (124,1.6), (124,1.4), (124,1.2), (124,1)}의 모델들이 탐색되었다. 이러한 탐색된 모델 중에서 테스트 오차가 최적인 영역에 속하는 모델들과 이들의 분류 성능은 아래 표 3과 같다.

상기 모델 중에서 테스트 오차가 가장 적은 경우는 커널의 개수(m)가 m=121, 124일 때이고, 이에 대응하는 분류 정확도는 99.1%이다. 이것은 분류 성능이 매우 높은 반면, 모델의 복잡도를 나타내는 커널의 개수가 다소 많음을 알 수 있다. 분류 성능이 높은 것은 Monks1이 비교적 쉬운 문제이기 때문인 것으로 판단되며, 데이터에 내재된 노이즈가 없기 때문에 주어진 데이터를 완전히 나타내기에 충분한 모델 복잡도가 높은 모델이 좋은 성능을 보임을 알 수 있다.

3 Monks2에 대한 실험 결과

Monks2에 관한 기존의 연구 결과물들을 살펴보면, 대부분의 경우 Monk1, Monks3에 비해 상대적으로 매우 높은 테스트 오차를 나타낸다[6,7,8,9]. 이것은 Monks2

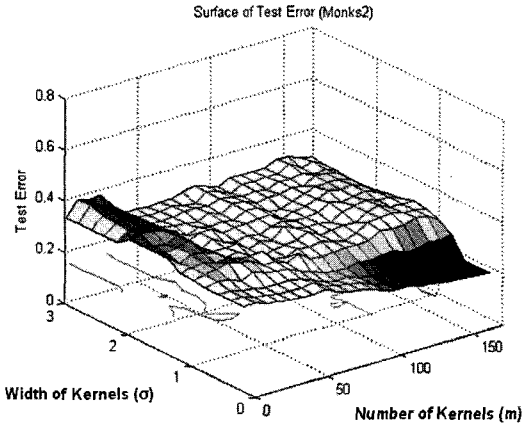


그림 4. Monks2에 대한 전체 커널 변수(m,σ) 공간에서 생성된 테스트 오차 표면

Fig. 4. Surfaces of test errors generated over kernel parameter (m,σ) space for Monks2.

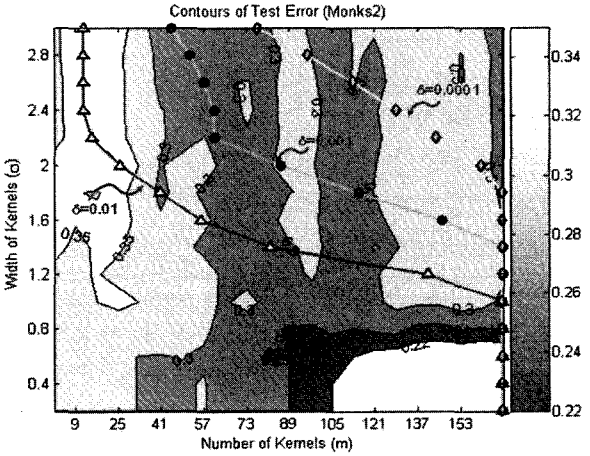


그림 5. Monks2에 대한 테스트 오차 컨투어 및 δ별 탐색 후보 모델의 예

Fig. 5. Contours of test errors and candidate models explored for different δ in Monks2.

가 매우 복잡하여 학습하기 어려운 문제로 알려진 것과 연관이 있다. 본 실험에서도 표 4에서와 같이 Monks1, Monks3에 비해 높은 테스트 오차를 나타내었다.

169개의 데이터 샘플로 구성된 Monks2 학습 데이터에 대하여 *m*=(1:8:169)와 σ=(0.2:0.2:3)의 범위에서 모든 가능한 (m,σ) 조합에 대한 RBF 모델을 생성하고 이에 대한 테스트 오차 표면 및 컨투어를 그림 4, 5와 같이 각각 생성하였다. 그림 5에서 흰색으로 나타나는 공간은 테스트 오차가 최적인 범위*를 나타낸다.

RBF 모델을 생성하기 위하여 *RC* 기준치로서 δ

* Monks2에 대하여 최근까지 발표된 결과 중 테스트 오차 최저치인 0.22 (즉, 오분류율이 22%)보다 적은 범위를 나타낸다.

표 4. Monks2에서 δ값에 따른 최적의 테스트 오차 범위 내의 모델들

Table 4. Selected models having their test errors in the optimal range for different δ in Monks2.

δ	σ	*m*	TestCE	Classification Accuracy
0.01	0.2	169	0.181	81.9%
	0.4	169	0.181	81.9%
	0.6	169	0.181	81.9%
0.001	0.2	169	0.181	81.9%
	0.4	169	0.181	81.9%
	0.6	169	0.181	81.9%
0.0001	0.2	169	0.181	81.9%
	0.4	169	0.181	81.9%
	0.6	169	0.181	81.9%

=0.01, 0.001, 0.0001의 세 가지 값을 사용하였고, 각 δ값에 대하여 알고리즘에 의해 생성되어 탐색된 모델들은 그림 5에 표기된 바와 같다. 구체적으로 δ=0.01인 경우, (m,σ)={(12,3), (12,2.8), (12,2.6), (12,2.4), (17,2.2), (27,2), (39,1.8), (53,1.6), (79,1.4), (121,1.2), (124,1)}에 대응하는 모델들이 탐색되었으며, δ=0.001인 경우에는 (m, σ)={(42,3), (48,2.8), (53,2.6), (57,2.4), (62,2.2), (81,2), (103,1.8), (122,1.6), (124,1.4), (124,1.2), (124,1)}이 탐색되었다. δ=0.0001인 경우에는 (m,σ)={(75,3), (89,2.8), (101,2.6), (112,2.4), (121,2.2), (124,2), (124,1.8), (124,1.6), (124,1.4), (124,1.2), (124,1)}의 모델들이 탐색되었다. Monks2에서 탐색된 모델 중 테스트 오차가 최적인 영역에 속하는 모델들과 이들의 분류 성능은 아래 표 4와 같다.

표 4에 나타난 바와 같이, 서로 다른 δ값을 사용한 경우에 있어서도 선택된 모든 모델에 대한 커널의 개수(m)는 m=169로서, 분류 정확도는 81.9%를 나타낸다. 이것은 Monks1에 비해 성능 면에서 매우 낮은 수치이며 문제의 복잡도가 다른 문제에 비해 상대적으로 높기 때문에 나타난 결과로 판단된다. 또한, 노이즈가 포함되어 있지 않기 때문에 주어진 데이터를 완전히 나타낼 수 있는 모델의 복잡도가 높은 모델이 좋은 성능을 보임을 알 수 있다.

4. Monks3에 대한 실험 결과

122개의 데이터 샘플로 구성된 Monks3 학습 데이터에 대하여 *m*=(1:6:122)와 σ=(0.2:0.2:3)의 범위에서 모든 가능한 (m,σ) 조합에 대한 RBF 모델을 생성하고 이에 대한 테스트 오차 표면 및 컨투어를 그림 6, 7과 같이 각각 생성하였다. 그림 7에서 흰색으로 나타나는 공간

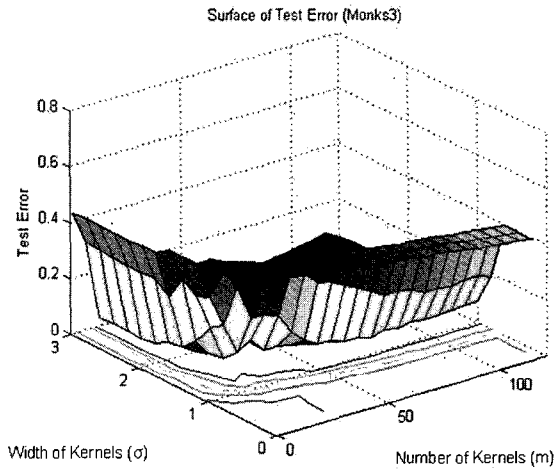


그림 6. Monks3에 대한 전체 커널 변수(m,σ) 공간에서 생성된 테스트 오차 표면

Fig. 6. Surfaces of test errors generated over kernel parameter (m,σ) space for Monks3.

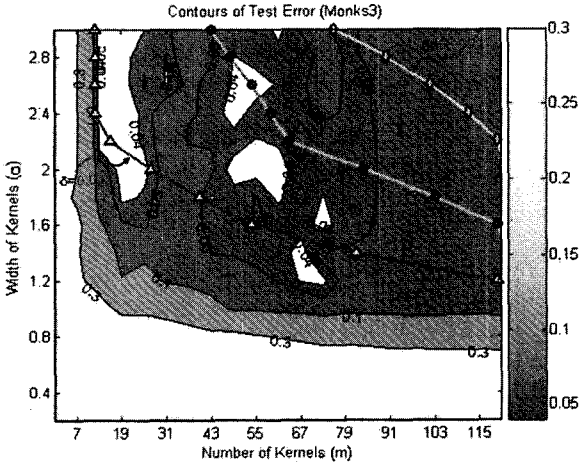


그림 7. Monks3에 대한 테스트 오차 컨투어 및 δ별 탐색 후보 모델의 예

Fig. 7. Contours of test errors and candidate models explored for different δ in Monks3.

은 테스트 오차가 최적인 범위*를 나타낸다. RBF 모델을 생성하기 위하여 *RC* 기준치로서 δ=0.01, 0.001, 0.0001의 세 가지 값을 사용하였고, 각 δ값에 대하여 알고리즘에 의해 탐색된 모델들은 그림 7에 표기된 바와 같다. 이러한 모델 중 테스트 오차가 최적인 영역에 속하는 모델들과 이들의 분류 성능은 아래 표 5와 같다.

상기 모델 중에서 테스트 오차가 가장 적은 경우는 커널의 개수(m)가 m=12일 때이고, 이에 대응하는 분류 정확도는 97.2%이다. 이것은 비슷한 수준의 Monks1 문

* Monks3에 대하여 최근까지 발표된 결과 중 테스트 오차 최저치인 0.04 (즉, 오분류율이 4%)보다 적은 범위를 나타낸다.

표 5. Monks3에서 δ값에 따른 최적의 테스트 오차 범위 내의 모델들

Table 5. Selected models having their test errors in the optimal range for different δ in Monks3.

δ	σ	m	TestCE	Classification Accuracy
0.01	1.6	54	0.039	96.1%
	2.2	16	0.035	96.5%
	2.4	12	0.028	97.2%
	2.6	12	0.028	97.2%
	2.8	12	0.028	97.2%
	3.0	12	0.028	97.2%
0.001	2.4	59	0.035	96.5%
0.0001	-	-	-	-

제에 비해, 분류 성능이 상대적으로 낮은 반면에 커널의 개수는 훨씬 적은 모델을 필요로 함을 알 수 있다. 분류 성능이 낮은 것은 Monks3 데이터에 내재된 노이즈 때문에 나타난 결과이며, 커널의 개수에 있어서도 주어진 데이터를 완전히 나타내는 모델 보다는 노이즈를 고려한 어느 정도의 오차를 허용하는 *RC* 기준치가 필요하기 때문에 모델 복잡도를 나타내는 커널의 개수가 급격히 줄어든 것으로 해석된다.

V. 결 론

본 논문에서는 데이터 마이닝에의 응용을 위한 가우시안 커널 RBF 모델의 유용성을 Monk's problem 분석을 통해 살펴보았다. 실제적으로 데이터 마이닝 모델을 개발해야 하는 경우, 주어진 상황에 따라 최적의 모델을 판단하는 기준이 달라질 수 있다. 예를 들어, 모델의 복잡도에 상관없이 분류 성능이 최고인 모델을 요하는 경우가 있는 반면, 분류 성능이 어느 정도 합당한 수준 안에서 모델 복잡도가 낮은 것이 중요한 기준이 될 수도 있다. 이렇게 주어진 상황에 따라 모델의 최적성(optimality)은 다양한 기준에서 고려될 수 있으며, 본 실험에서 사용된 모델링 방법은 이러한 상황을 모델링 과정에서 적절히 반영할 수 있는 기반을 제공하고 있다.

Monk's problems에 대해 본 실험에서 생성된 분류 성능 면에서의 최적의 모델 선정 결과는 아래 표 6에서와 같다. 표 6에 나타난 바와 같이, Monks1, Monks2, Monks3에 대한 최고의 분류 성능은 테스트 데이터에 대하여 각각 99.1%, 81.9%, 97.2%의 분류 정확도를 나타내었고, 이 때의 학습 오차는 각각 0.0%, 0% 6.6%을 나타내었다.

표 6. Monk's Problem에 대한 분류 성능 면에서의 최적 모델 선정 결과 및 분류 정확도

Table 6. Our best models in terms of classification performance for Monk's problems and their classification accuracy.

Dataset	σ	m	TrainCE (Accuracy)	TestCE (Accuracy)
Monks1	2.2	121	0.0 (100%)	0.009 (99.1%)
Monks2	0.2-0.6	169	0.0 (100%)	0.181 (81.9%)
Monks3	2.4-3.0	12	0.066 (93.4%)	0.028 (97.2%)

표 7. Monk's Problem에 대한 모델 복잡도 면에서의 최적 모델 선정 결과 및 분류 정확도

Table 7. Our best models in terms of model complexity for Monk's problems and their classification accuracy.

Dataset	σ	m	TrainCE (Accuracy)	TestCE (Accuracy)
Monks1	2.8	89	0.0 (100%)	0.021 (97.9%)
Monks2	0.2-0.6	169	0.0 (100%)	0.181 (81.9%)
Monks3	2.4-3.0	12	0.066 (93.4%)	0.028 (97.2%)

한편, 표 7에서는 최적 영역에 속한 분류 성능을 나타내면서 모델의 복잡도 면에서 최적의 모델로 선정된 결과를 보여주고 있다. 특히, Monks1의 경우 커널의 개수가 표 6의 121에서 89개로 크게 감소한 반면 분류 정확도는 99.1%에서 97.9%로 약간 낮아졌음을 알 수 있다.

또한, 앞의 테스트 오차 표면들(그림 2, 4, 6)에서도 나타나듯이 표 6, 7에서 제시된 모델이외에도 그 주변의 다소 넓은 범위의 커널의 수(m)과 커널 폭(σ)에 대하여도 유사한 성능을 나타냄을 알 수 있다. 따라서 모델 복잡도나 분류 성능 등 주어진 상황에 따라 모델의 최적성 판단 기준을 적절히 설정함으로써 실제 상황에 가장 적합한 최적의 모델을 융통성 있게 선택할 수 있다.

이러한 *RC* 기반 RBF 모델링 방법은 다소 임의적(arbitrariness)인 경향이 있는 일반적인 많은 다른 모델링 방법들에 비해, 사용이 용이할 뿐만 아니라 주어진 환경에 따라 최적의 모델을 유연하게 결정할 수 있다는 특징 때문에 다양한 마이닝 환경에서 효율적으로 이용될 수 있으리라 판단된다.

참 고 문 헌

[1] L. Breiman, J.H. Friedman, R.A. Olshen and C.J. Stone, Classification and Regression Trees, Wadsworth, 1984.

[2] G.V. Kass, "An Exploratory technique for investigating large quantities of categorical data," Applied Statistics, pp.119-127, 1980.

[3] D. Michie, D.J. Spiegelhalter and C.C. Taylor (eds), Machine learning, Neural and Statistical Classification, Ellis Horwood, 1994.

[4] Q.Yang and X. Wu, "10 Challenging problems in data mining research," in presentation slides of IEEE conference on Data Mining, 15. Dec, 2005.

[5] S.B. Thrun et al, "The Monk's problems: a performance comparison of different learning algorithms," Technical Report CMU-CS-91-197, Carnegie Mellon University. 1991.

[6] H. Xiong, M.N. S. Swamy, and M. Omair Ahmad, "Optimizing the kernel in the empirical feature space," IEEE Transactions on Neural Networks. March 2005.

[7] M. W. Mitchell, "An architecture for situated learning agents," Ph.D. Dissertation, Monash University, Australia, 2003.

[8] M. Casey and K. Ahmad, "In-situ learning in multi-net systems," Lecture Notes in Computer Science, vol. 3177, pp. 752-757, 2004.

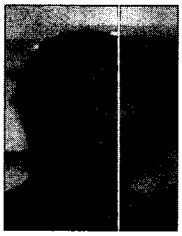
[9] K. Toh, Q-L Tran and O. Srinivasan, "Benchmarking a reduced multivariate polynomial pattern classifier," IEEE Trans. on Pattern Anal. and Machine Intelligence, vol.16, no.2, pp. 460-474, 2005.

[10] S. H. Huang, "Dimensionality reduction in automatic knowledge acquisition: a simple greedy search approach," IEEE Transactions on Knowledge and Data Engineering. vol. 16, no. 6, pp. 1364-1373, 2003.

[11] S. Saxon and Alwyn Barry, "XCS and the Monk's problems in learning classifier systems: from foundations to applications," P.L. Lanzi et al, Ed., Lecture Notes in Computer Science, vol. 1813, pp. 440-448, 2000.

[12] A. L. Goel and Miyoung Shin, "Radial basis functions: an algebraic approach (with data mining applications)," Tutorial notes in European conference on Machine Learning, Pisa, Italy, September 2004.

저 자 소 개



신 미 영(정회원)

1991년 연세대학교 전산과학과 학사 졸업.

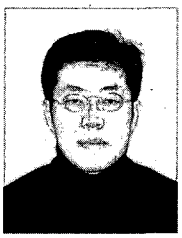
1993년 연세대학교 전산과학과 석사 졸업.

1998년 미국 Syracuse Univ., EECS dept. (Ph.D)

1999년~2005년 3월 한국전자통신연구원 선임 연구원

2005년 4월~현재 경북대학교 전자전기컴퓨터학부 조교수

<주관심분야 : 패턴인식, 바이오인포매틱스>



박 준 구(정회원)

1994년 서울대학교 제어계측공학 학사 졸업.

1996년 서울대학교 제어계측공학 석사 졸업.

2001년 서울대학교 전기컴퓨터 공학부 박사 졸업.

2001년~2005년 3월 삼성 정보통신연구소 책임연구원

2005년 4월~현재 경북대학교 전자전기컴퓨터학부 조교수

<주관심분야 : 이동통신, 모바일 네비게이션>